

Copyright is owned by the Author of the thesis. Permission is given for a copy to be downloaded by an individual for the purpose of research and private study only. The thesis may not be reproduced elsewhere without the permission of the Author.

NESTED TANDEM REPEAT COMPUTATION AND ANALYSIS

A thesis presented in partial fulfilment
of the requirements for the degree of

Doctor of Philosophy
in Computational Biology

at Massey University

Atheer Matroud
2013

Copyright © 2013 by Atheer Matroud

Abstract

Biological sequences have long been known to contain many classes of *repeats*. The most studied repetitive structure is the *tandem repeat* where many approximate copies of a common segment (the *motif*) appear consecutively. In this thesis, a complex repetitive structure is investigated. This repetitive structure is called a *nested tandem repeat*. It consists of many approximate copies of two motifs interspersed with one another.

This thesis is a collection of published and in progress papers. Each paper addresses a computational problem related to the analysis of nested tandem repeats. Nested tandem repeats have been observed in the intergenic spacer of the ribosomal DNA gene in *Colocasia esculenta*. The question of whether such repeats can be found elsewhere in biological sequence databases is addressed and `NTRFinder`, a software tool to detect nested tandem repeats, is described. Another problem that arises after detecting a nested tandem repeat is the alignment of the nested tandem repeat region against its two motifs. An algorithm that guarantees an optimal solution to this problem is introduced. After detecting nested tandem repeats and identifying their structures, the identification of the motif boundaries is an unsolved problem which arises not only in nested tandem repeats but in tandem repeats as well. Heuristic solutions to this problem are implemented and tested. In order to compare two tandem repeat sequences an algorithm that aligns a hypothetical ancestral sequence of both sequences against each sequence is presented. This algorithm considers substitutions, deletions, and unidirectional duplication, namely, from ancestor to descendant.

Acknowledgements

I am sincerely thankful to my supervisors Prof. Mike Hendy and Dr. Chris Tuffley for their continuous support from day one through my candidature. It has been a privilege to work with both of them. Without their guidance and patience this work would have not been possible. I am forever grateful for their continuous support. I would also like to thank A/Prof. David Bryant for his support and help in my studies after I moved to Dunedin.

I would like to thank the Allan Wilson Centre for Molecular Ecology and Evolution for funding my study. I would also like to thank the Institute of Fundamental Sciences for the financial support they have provided to me.

I would like to thank Prof. Jens Stoye for the fruitful discussions I had with him during my visit to his lab and during our meeting in Campo Grande. Many thanks to A/Prof. Nadia El-Mabrouk for her support of my visit to her lab and for her advice. I would also like to thank A/Prof. Eric Rivals for his suggestions and comments.

I am grateful to Prof. Hamish Spencer for giving me his office to use. I am also thankful to the Department of Zoology staff for their hospitality. I am thankful to the Mathematics and Statistics department members at the University of Otago for hosting me and being so friendly.

I would like to thank all my colleagues in my office at Massey University and all colleagues in the “boffin lounge” for their company, help and support.

Last but not least, I am very grateful to my wife Zainab for her patience and strong support throughout my study. I am thankful to my children, Qaswar, Jaafar, and Durar for being such a great inspiration to me. I am grateful to my father, my mother, Naseer, Abeer, Akeel, and Aseel for their continued support.

Contents

Abstract	iii
Acknowledgements	v
1 Introduction	1
1.1 Candidate's Note	2
1.2 Definitions	2
1.2.1 Sequences, edit operations and the edit distance	2
1.2.2 Classification of Tandem Repeats	3
1.2.3 Nested Tandem Repeats	4
1.3 A duplication model for tandem repeats and nested tandem repeats	6
1.4 Overview	8
2 Literature Review	11
2.1 Motivation	11
2.2 Models of tandem repeat evolution	12
2.3 Detection of tandem repeats	13
2.4 Alignment	15
2.5 Alignment of two tandem repeat sequences	15
3 Observations on the nested tandem repeat found in taro	17
3.1 Nested tandem repeat structures in NZ1 and JP1	18
3.2 Nested tandem repeat variants sequence	18
3.3 Variants graph	18
3.4 Expected number of parallel substitutions	23
3.5 Variants frequency distribution	24
3.6 Variants spread	25

4	NTRFinder: A Software Tool to Find Nested Tandem Repeats	27
4.1	Abstract	27
4.2	Introduction	27
4.3	Material and Methods	28
4.4	Results	31
4.4.1	Tests on simulated data	31
4.4.2	Tests on real sequence data	32
4.4.3	More complex structures	32
4.4.4	Running time	32
4.5	Discussion	33
4.6	Conclusion	33
5	An algorithm to solve the motif alignment problem for approximate nested tandem repeats in biological sequences	39
5.1	Abstract	39
5.2	Introduction	40
5.3	Definitions	41
5.3.1	Alphabets and strings	41
5.3.2	The edit distance	41
5.3.3	Tandem repeats and nested tandem repeats	42
5.3.4	Alignment	43
5.4	The motif alignment problem for approximate nested tandem repeats . . .	44
5.4.1	The problem	44
5.4.2	Solution to the problem via nested wrap-around dynamic programming	44
5.4.3	Correctness of the algorithm	46
5.4.4	Extension to nested tandem repeats with three or more motifs . .	49
5.5	Conclusion	50
6	A comparison of three heuristic methods for solving the parsing problem for tandem repeats	51
6.1	Abstract	51
6.2	Introduction	52

6.3	Definitions and Background	54
6.4	The importance of the parsing problem	55
6.5	Heuristic methods to estimate tandem repeat parsing	56
6.5.1	PAIR — the adjacent pairs method	58
6.5.2	VAR — the number of variants method	59
6.5.3	MST — the minimum spanning tree method	59
6.6	Results and discussion	60
6.6.1	Tests on simulated data	60
6.6.2	Tests on real sequence data	65
6.7	Conclusion	66
7	Ancestor-descendant alignment of tandemly repeated sequences	71
7.1	TR maps	72
7.2	Edit operations and edit distance	73
7.3	Ancestor-descendant repeat distance	74
7.3.1	The ancestor-descendant alignment problem for (N)TR sequences	74
7.3.2	Solution to the ancestor-descendant alignment problem	76
7.3.3	Correctness of the algorithm	77
7.4	A Longest Common Subsequence approach to estimating the most recent common ancestor	83
7.5	An application to real DNA sequences	83
7.6	Conclusion	84
8	Conclusion	87
8.1	Future work	88
A	Published chapters	91